**Local n8n AI Agents in 5 Minutes (FREE and no code)**

**Transcrito por** [**TurboScribe.ai**](https://turboscribe.ai/pt/?ref=docx_export_upsell)**.** [**Atualize para Ilimitado**](https://turboscribe.ai/pt/subscribed?ref=docx_export_upsell) **para remover esta mensagem.**

Hoje eu vou mostrar como você pode usar seus agentes de A.I. N8n completamente gratuitos. Há um monte de diferentes modelos open-source que podemos instalar e usar e eu vou mostrar a vocês o quão simples é a configuração. Tudo o que você tem que fazer é seguir este vídeo.

Então vamos começar. Então, como prerequisito, eu estou assumindo que você já instalou sua N8n localizada usando o A.I. Self-hosted Starter Kit. Se você ainda não fez isso, veja este vídeo que eu vou botar aqui em cima em breve e depois pule de volta para este vídeo.

Então, uma vez que você tem tudo instalado, você pode ver que o que temos agora é uma versão localizada de N8n funcionando. Temos o nosso Ollama, que é um modelo localizado, mas aqui tudo o que podemos ver é que temos o Ollama 3.2 mais recente. Então o que precisamos fazer é basicamente ir ao Ollama e escolher de todos os diferentes modelos open-source que estão disponíveis e apenas instalá-los na nossa máquina, bem rápido.

E tudo isso vai ser gratuito. Então deixe-me mostrar a vocês como fazemos isso. Então você vai para ollama.com, eu vou colocar o link na descrição e você nem precisa vir aqui e acessar nada.

Apenas vá aqui, clique em modelos e agora temos todos os diferentes modelos que podemos acessar e começar a usar com nossos agentes de A.I. N8n. Então temos DeepSeek, que foi o modelo mais popular, 48.1 milhões de pulos. Temos todos esses outros que escolhemos, temos alguns mais modelos Ollama e este é basicamente o nosso repositório de diferentes modelos de A.I. que podemos testar grátis.

E só para mostrar a vocês um exemplo, vamos fazer DeepSeek R1, que é um modelo de pensamento e tudo o que vamos fazer é clicar nele e agora temos algumas coisas para olhar. Então, primeiro de tudo, temos os diferentes parâmetros que a versão open-source de DeepSeek é oferecida. Então temos 1.5 bilhões, 7 bilhões, 8 bilhões, todo esse tipo de coisa.

E aqui em baixo você pode ver todos os diferentes modelos, com seus tamanhos e suas janelas de contato. E eu não sou realmente um experto em hardware e tudo, mas o que eu faria é ir para a perplexidade ou o chat e dizer Hey, essas são as especificações do meu computador, aqui está meu RAM, meu VRAM, meu poder de processamento, todo esse tipo de coisa, e eu diria, qual modelo eu deveria ser capaz de usar? E só para referência, esses são bem pequenos, se a gente vai aqui para a minha perplexidade de pesquisa ou se a gente olha para alguns dos modelos mais populares de open-source, como GPT e Claude, GPT-4 é quase 1.8 trilhões, o que é bem enorme. Você sabe, como GPT-4.0 é 200 bilhões, Claude 3.5 SONNET é em torno de 175 bilhões.

Então esses modelos de aberto-source serão muito mais rápidos e mais poderosos do que a maioria desses modelos de aberto-source. Essa gapa está acabando devagar e um dia haverá modelos de aberto-source que serão tão poderosos quanto esses. Mas agora, essa é a limitação de modelos de aberto-source.

E você também vai ser limitado baseado na verdadeira máquina local que você vai instalar e usar esses modelos. Mas de qualquer forma, agora que isso está fora do caminho, vamos, por exemplo, neste caso, baixar DeepSeek R1 que tem 1.5 bilhão de parâmetros. Então tudo o que eu vou fazer é ir e clicar nesse botão de copia aqui.

Então eu estou apenas copiando o nome DeepSeek-R1, colon, 1.5b. Eu vou abrir o nosso docker e podemos ver o nosso conteineiro com todos esses diferentes conteineiros neles. E o que eu vou fazer é basicamente abrir o conteineiro Ollama que tem a imagem do Ollama nele. E tudo o que eu preciso fazer é clicar no Exec e eu vou apenas escrever Ollama pull.

Eu vou apenas escrever o que a gente apenas copiou, que foi DeepSeek-R1. Eu vou clicar em Enter. Agora o que está fazendo é que está puxando DeepSeek.

Está instalando na nossa máquina. E então, uma vez isso terminado, vamos poder acessar isso no end-to-end. Ok, então a gente acabou de receber nossa mensagem de sucesso.

Eu vou voltar para o nosso local end-to-end e abrir esse nodo Ollama. E agora, se eu clicar nesse arco, vemos DeepSeek-R1 aqui. Então eu trocei, agora estamos usando DeepSeek.

E vamos testar que funciona. Eu vou apenas dizer, me diga uma brincadeira. E agora não estamos mais acessando nosso local modelo Ollama.

Estamos usando nosso local DeepSeek-R1. E você pode ver, porque é um modelo de pensamento, e ele disse, Pense, pense, por que os esqueletos não lutam um com o outro? Porque eles não têm os ossos. E então, é realmente tão simples.

E agora, tudo o que a gente tem que fazer é voltar para Ollama, clicar em Modelos, e a gente pode olhar todos os diferentes modelos de AI, assim como os diferentes parâmetros que são oferecidos, e apenas puxá-los para o nosso docker. Então, por exemplo, digamos que a gente quisesse puxar em Quen. E eu quero puxar em Quen 1.7 bilhões de parâmetros.

Eu só copiaria o nome. Vou para o nosso Ollama no docker, no tab Exec, vamos dizer Ollama, puxar, e vou apagar esse valor, e agora vamos puxar em Quen, assim, super simples. Temos nossa mensagem de sucesso, vou voltar para o end-to-end, vou clicar em Ollama, e abrimos, e podemos ver que agora temos o nosso Quen 3, 1.7 bilhões de parâmetros, modelo.

E isso é tão legal, porque, de novo, isso está funcionando localmente, no meu PC lá embaixo, tudo vai ser completamente livre, e temos tantas opções diferentes para escolher. E, como eu disse, eu não sou um experto em modelos, pesadelos e parâmetros, mas, só tenha em mente, eu provavelmente não iria depender muito de um 1.7 bilhões de parâmetros ou 1.5 bilhões de parâmetros para qualquer coisa além de uso muito básico de conversa. Se você quiser ter agentes com chamadas de ferramentas e todo esse tipo de coisa usando modelos locais, você provavelmente vai querer pelo menos, tipo, 14 bilhões ou 32 bilhões.

Mas, de novo, eles são completamente livres, então você pode testar o quanto quiser. De qualquer forma, deixe-me mostrar para vocês alguns comandos ajudantes que você pode usar dentro desse Exec do Ollama Container. Então, digamos que você só quer ver quais modelos você tem disponível, você pode ir e escrever Ollama List.

E isso vai nos mostrar, ok, você tem Quen, DeepSeq e Ollama 3.2. Agora, digamos que nós querermos se livrar de um. Vamos só se livrar do Quen que a gente apenas instalou. Eu só iria escrever Ollama RM, que significa para remover, e eu só iria poder escrever o nome do modelo que nós queremos remover, clicar em Enter, e agora diz que é o Quen deletado.

Se eu voltar para os nossos modelos locais, abrimos esse cara, nós não mais temos Quen. Como você pode ver, nós só temos DeepSeq ou Ollama para escolher. E é tão simples.

Ele atualiza em tempo real, você não precisa nem, você sabe, docarcompilar e virar de novo. Nós basicamente estamos prontos. Tão rápido e fácil.

E, como eu disse, você tem todos esses modelos para escolher. Então, isso é tudo por esse vídeo. Eu sei que foi um vídeo rápido, mas eu só queria mostrar para vocês o quão simples e fácil é conseguir esses diferentes modelos locais funcionando, e isso pode parecer um pouco intimidante no começo, mas depois desse vídeo, você pode ver que não é tão ruim de fato.

Então, se você aprendeu algo novo ou apreciou esse vídeo, dê um like, isso definitivamente me ajuda um monte, e eu vejo vocês todos no próximo. Obrigado, pessoal.

**Transcrito por** [**TurboScribe.ai**](https://turboscribe.ai/pt/?ref=docx_export_upsell)**.** [**Atualize para Ilimitado**](https://turboscribe.ai/pt/subscribed?ref=docx_export_upsell) **para remover esta mensagem.**